

BAB I

PENDAHULUAN

1.1 Latar Belakang

Regresi logistik biner adalah metode analisis statistik yang digunakan untuk memodelkan hubungan antara satu atau lebih variabel prediktor dengan variabel respon yang bersifat biner atau dikotomis. Variabel respon dalam regresi logistik ini hanya memiliki dua kemungkinan hasil, seperti berat badan lahir rendah (BBLR) atau tidak BBLR (Anisa et al., 2023). Selain itu, regresi logistik biner memungkinkan penggunaan berbagai metode evaluasi performa model, seperti akurasi, sensitivitas, dan spesifisitas yang sangat penting dalam klasifikasi biner (Hosmer et al., 2013).

Dalam berbagai penelitian regresi logistik biner banyak digunakan karena kesederhanaannya, interpretasi yang mudah, serta kemampuannya untuk bekerja dengan data yang memiliki hubungan linier antara prediktor dan logit dari variabel respon. Namun, regresi logistik biner memiliki kelemahan ketika diterapkan pada data yang tidak seimbang, yaitu ketika jumlah data dari salah satu kelas jauh lebih banyak dibandingkan kelas lainnya. Ketika model diterapkan pada data, hasil model cenderung memiliki bias terhadap kelas mayoritas, sehingga performa model terhadap kelas minoritas, yang sering kali lebih penting, menjadi buruk (Fernandez et al., 2018). Menurut Li & Sun dalam penelitian Mustaqim et al. (2019), jika proporsi kelas minoritas kurang dari 35% dari total data, maka dikategorikan sebagai data yang tidak seimbang.

Salah satu solusi untuk mengatasi masalah ketidakseimbangan data adalah penggunaan *Synthetic Minority Oversampling Technique* (SMOTE). SMOTE bekerja dengan menghasilkan sampel sintesis baru pada kelas minoritas melalui interpolasi linier antar titik data minoritas yang ada, sehingga dapat meningkatkan keseimbangan antara kelas mayoritas dan minoritas (Chawla et al., 2002). Meskipun SMOTE telah terbukti efektif dalam berbagai aplikasi, teknik ini memiliki keterbatasan ketika diterapkan pada dataset yang berisi variabel nominal dan kontinu. Hal tersebut disebabkan oleh fakta bahwa interpolasi linier, yang merupakan inti dari SMOTE, hanya dapat diterapkan pada variabel kontinu, sementara variabel nominal tidak memiliki hubungan numerik yang memungkinkan interpolasi (Blagus & Lusa, 2013).

Sebagai solusi untuk keterbatasan SMOTE, dikembangkanlah metode *Synthetic Minority Oversampling Technique - Nominal Continuous* (SMOTE-NC), yang dirancang untuk menangani kumpulan data yang memiliki campuran variabel nominal dan kontinu. Variabel nominal adalah jenis variabel kualitatif yang terdiri dari kategori-kategori yang tidak memiliki nilai numerik dan tidak memiliki urutan atau ranking intrinsik. Variabel nominal sering menggunakan kategori yang bersifat biner atau dikotomi, seperti status berat badan lahir bayi (BBLR atau tidak BBLR), kadar Hb (Normal atau anemia), ibu mengalami komplikasi kehamilan (Ya atau Tidak) dan status pekerjaan ibu (Bekerja atau Tidak) (Dettori & Norvell, 2018). Sedangkan, variabel kontinu adalah jenis variabel yang dapat mengambil nilai dalam rentang

yang tidak terbatas. Dalam konteks penelitian, variabel ini penting untuk analisis statistik, seperti regresi yang memerlukan data kontinu untuk menghasilkan estimasi yang akurat (Hadjar, 2018). Dalam variabel kontinu jumlah terukur dapat mengasumsikan nilai apapun dalam kisaran tertentu, seperti umur kehamilan, paritas dan usia ibu (Sauerbrei and Royston, 2010). Salah satu jenisnya adalah data dengan skala rasio, yang dicirikan oleh titik nol sejati sehingga memungkinkan perbandingan besaran absolut (Kim and Frisby, 2019).

Dalam SMOTE-NC, interpolasi dilakukan pada variabel kontinu, sedangkan variabel nominal diperlakukan dengan memilih nilai berdasarkan probabilitas kemunculannya di kelas minoritas. Teknik tersebut memastikan bahwa data sintetis yang dihasilkan lebih relevan secara statistik dan menghasilkan nilai yang sesuai untuk variabel nominal. Penerapan SMOTE-NC telah menunjukkan peningkatan kinerja klasifikasi pada data yang tidak seimbang di berbagai bidang, terutama dalam kumpulan data yang melibatkan variabel nominal dan kontinu secara bersamaan (Mukherjee et al, 2021).

Metode ini bekerja dengan cara mencari k sampel terdekat untuk setiap data dalam minoritas, setelah itu dibuatkan duplikasi sintetis data yang diinginkan secara persentase antara data minoritas dengan *k-nearest neighbors* (K-NN) yang dipilih secara acak (Ratih et al., 2022). Pada versi standar K-NN semua tetangga memiliki bobot yang sama. Namun, pada *weighted k-nearest neighbors* (WK-NN) tetangga diberi bobot berdasarkan jaraknya (Oren et al, 2016). Metode WK-NN memiliki kelebihan dibanding K-NN standar, terutama untuk menangani masalah jumlah kelas yang tidak seimbang dalam pengklasifikasian data (Aisha et al, 2024).

Pemodelan regresi logistik biner untuk data tidak seimbang melalui kombinasi SMOTE-NC dengan WK-NN. Adapun penelitian-penelitian sebelumnya yang telah menggunakan metode-metode ini diantaranya (Ratih et al., 2022) menemukan bahwa SMOTE-NC dapat meningkatkan akurasi 85,29%, sensitivitas 91%, dan presisi 94%. Hasil tersebut menunjukkan bahwa SMOTE-NC dapat menangani ketidakseimbangan data (puas dan tidak puas terhadap pelayanan mahasiswa) dengan lebih baik, menghasilkan klasifikasi yang lebih akurat. sementara penelitian oleh (Islahulhaq et al., 2021) dengan menerapkan regresi logistik biner bersama dengan metode SMOTE-NC, penelitian tersebut berhasil meningkatkan akurasi prediksi dan mengatasi masalah ketidakseimbangan pada kelas pembiayaan baik dan pembiayaan buruk/macet.

Dalam konteks penelitian ini, pemodelan regresi logistik biner melalui kombinasi SMOTE-NC dengan WK-NN untuk menangani masalah ketidakseimbangan data pada prediksi Berat Badan Lahir Rendah (BBLR). Kelahiran bayi dengan BBLR saat ini merupakan masalah penting pada bidang reproduksi manusia. Kelahiran bayi BBLR merupakan penyumbang besar pada kesakitan dan kematian neonatal jangka pendek maupun jangka panjang. BBLR memiliki hubungan yang signifikan terhadap kematian neonatal di Indonesia, dengan BBLR menjadi faktor yang paling signifikan (Rostina et al., 2022)

Di Indonesia sendiri khususnya di Provinsi Sulawesi Tenggara kasus BBLR cukup memprihatinkan karena dalam beberapa tahun terakhir terjadi kenaikan

jumlah kasus. Pada tahun 2020 terdapat 1.294 kasus, meningkat menjadi 1.572 kasus pada tahun 2021. Sementara itu, jumlah BBLR di Kabupaten Buton Tengah pada tahun 2020 terdapat 73 kasus (3,40%) dari 2.146 bayi lahir meningkat pada tahun 2021 menjadi 131 kasus (5,61%) dari 2.333 bayi lahir (BPS Sultra, 2020; 2021). masalah BBLR ini menjadi tantangan kesehatan yang serius, terutama di daerah dengan akses pelayanan kesehatan yang terbatas. Ketidakseimbangan dalam data kesehatan ibu dan anak sering terjadi karena lebih banyak data tersedia dari kasus kelahiran bayi dengan berat badan normal (tidak BBLR), sementara jumlah data tentang kasus BBLR sedikit. Masalah ini memperkuat kebutuhan akan penerapan teknik pengolahan data yang lebih canggih seperti kombinasi SMOTE-NC dengan WK-NN.

Dalam data kesehatan ibu dan bayi, sering kali ditemukan ketidakseimbangan kelas yang signifikan, yaitu jumlah bayi dengan BBLR jauh lebih sedikit dibandingkan dengan bayi yang lahir dengan berat badan normal (tidak BBLR). Sebagai contoh, variabel-variabel seperti usia ibu, status pekerjaan ibu, pendidikan, dan komplikasi biasanya mencakup variabel kontinu dan nominal yang saling berhubungan dalam mempengaruhi probabilitas kelahiran bayi dengan BBLR. Dengan adanya ketidakseimbangan data tersebut menyebabkan estimasi parameter yang bias dan kinerja klasifikasi yang buruk dalam model regresi logistik biner (Rahman et al, 2020).

Penggunaan kombinasi SMOTE-NC dengan WK-NN dalam penelitian ini memungkinkan penanganan data yang lebih baik dengan mempertimbangkan perbedaan antara variabel nominal dan kontinu. Misalnya, kombinasi SMOTE-NC dengan WK-NN akan menghasilkan sampel sintetis untuk variabel kontinu seperti umur kehamilan, paritas dan usia ibu melalui interpolasi linier, sementara nilai variabel nominal seperti Komplikasi Kehamilan, kadar Hb, dan Status Pekerjaan Ibu akan dipilih berdasarkan probabilitas kemunculan di kelas minoritas. Dengan demikian, model regresi logistik biner yang dilatih pada data yang telah diimbangi dengan SMOTE-NC diharapkan dapat meningkatkan akurasi prediksi pada kelas minoritas (BBLR), tanpa mengorbankan validitas data asli.

Berdasarkan latar belakang tersebut, penelitian ini penting untuk pemodelan regresi logistik biner kombinasi SMOTE-NC dengan WK-NN dalam memodelkan dan meningkatkan performa klasifikasi pada dataset BBLR di Kabupaten Buton Tengah yang memiliki ketidakseimbangan kelas. Dengan demikian, penulis tertarik melakukan penelitian dengan judul "Pemodelan Regresi Logistik Biner untuk Data Tidak Seimbang Melalui Kombinasi *Synthetic Minority Oversampling Technique - Nominal Continuous* dengan *weighted k-nearest neighbors*".

1.2 Rumusan Masalah

Berikut rumusan masalah berdasarkan latar belakang yang telah diuraikan:

1. Bagaimana hasil kombinasi SMOTE-NC dengan WK-NN dalam menangani data tidak seimbang pada BBLR di Kabupaten Buton Tengah?
2. Bagaimana bentuk estimasi parameter model regresi logistik biner melalui kombinasi SMOTE-NC dengan WK-NN?
3. Bagaimana hasil klasifikasi model regresi logistik biner melalui kombinasi SMOTE-NC dengan WK-NN dalam menangani data tidak seimbang pada BBLR di Kabupaten Buton Tengah?

1.3 Tujuan Penelitian

Berikut tujuan penelitian berdasarkan rumusan masalah yang telah dijelaskan:

1. Mendapatkan hasil kombinasi SMOTE-NC dengan WK-NN dalam menangani data tidak seimbang pada BBLR di Kabupaten Buton Tengah.
2. Mendapatkan bentuk estimasi parameter model regresi logistik biner melalui kombinasi SMOTE-NC dengan WK-NN.
3. Mendapatkan hasil klasifikasi model regresi logistik biner melalui kombinasi SMOTE-NC dengan WK-NN dalam menangani data tidak seimbang untuk memberikan akurasi terbaik pada BBLR di Kabupaten Buton Tengah.

1.4 Manfaat Penelitian

Setelah tercapainya tujuan penelitian ini, maka diharapkan:

1. Hasil bentuk estimasi parameter dan klasifikasi model regresi logistik biner melalui kombinasi SMOTE-NC dengan WK-NN dalam menangani data tidak seimbang dapat menjadi bahan referensi dalam hal program penanganan kasus BBLR di Kabupaten Buton Tengah.
2. Hasil bentuk estimasi parameter dan model regresi logistik biner melalui kombinasi SMOTE-NC dengan WK-NN dalam menangani data tidak seimbang dapat menjadi bahan kajian dalam pengembangan metode dalam ilmu statistika.

1.5 Batasan Masalah

Penelitian ini hanya menggunakan data rekam medik dari Puskesmas di Kecamatan Mawasangka dan Kecamatan Gu, Kabupaten Buton Tengah tahun 2022. Fokus utama penelitian adalah mengestimasi model dan meningkatkan performa klasifikasi regresi logistik biner pada data tidak seimbang. Upaya peningkatan dilakukan melalui kombinasi metode SMOTE-NC dengan WK-NN. Proses oversampling dilakukan sebesar 600% dengan jumlah tetangga terdekat sebanyak $K=5$.

1.6 Teori-Teori

Sub-sub teori ini membahas mengenai beberapa teori yang digunakan untuk membahas pokok pembahasan terkait regresi logistik biner dengan kombinasi SMOTE-NC) dengan WK-NN pada data Berat Badan Lahir Rendah (BBLR) sebagai berikut:

1.6.1 Uji Independensi

Uji independensi atau *Chi-square* mengevaluasi hubungan antara variabel nominal, membantu mengidentifikasi asosiasi yang signifikan. Dengan menghilangkan variabel yang tidak signifikan. Pengujian ini menguji dua kelompok data baik variabel prediktor yang berbentuk kategorik atau dapat juga dikatakan sebagai uji proporsi untuk dua peristiwa atau lebih (Mary, 2013).

Hipotesis:

H_0 : Tidak ada hubungan variabel respon (Y) dengan variabel prediktor (X).

H_1 : Ada hubungan variabel respon (Y) dengan variabel prediktor (X).

Statistik Uji:

Statistika uji yang digunakan ditampilkan pada persamaan berikut:

$$\chi^2 = \sum_{i=1}^n \sum_{j=1}^p \frac{(n_{ij} - e_{ij})^2}{e_{ij}} \text{ dengan } e_{ij} = \frac{n_i \times n_j}{n} \quad (1)$$

keterangan:

n_{ij} : Nilai observasi atau pengamatan pada baris ke- i kolom ke- j

e_{ij} : Nilai ekspektasi baris ke- i kolom ke- j

n_i : Nilai observasi pada kolom ke- i

n_j : Nilai observasi pada baris ke- j

n : Jumlah seluruh pengamatan

Daerah Kritis:

Tolak H_0 jika $\chi_{hitung}^2 > \chi_{(df, \alpha)}^2$ dengan derajat bebas $(n-1)$ ($p-1$)

1.6.2 SMOTE

SMOTE adalah metode *oversampling* yang menghasilkan data sintetis untuk meningkatkan representasi kelas minoritas. Berbeda dengan metode oversampling tradisional yang hanya menggandakan atau menyalin sampel minoritas yang ada, SMOTE menghasilkan sampel baru dengan cara melakukan interpolasi antara sampel minoritas yang dipilih dan beberapa tetangga terdekatnya (Blagus and Lusa, 2013). Metode ini diterapkan untuk menangani ketidakseimbangan jumlah sampel dalam suatu kelas. Teknik ini mensintesis sampel baru dari kelas minoritas untuk menyeimbangkan dataset dengan melakukan resampling sampel kelas minoritas atau membuat replikasi data minoritas. Replikasi ini dikenal sebagai data sintetis. Metode SMOTE bekerja dengan cara mencari k sampel terdekat untuk setiap data dalam minoritas, setelah itu dibuatkan duplikasi sintetis data yang diinginkan secara persentase antara data minoritas dengan k -nearest neighbors yang dipilih secara acak (Ratih et al., 2022).

1.6.3 SMOTE-NC

SMOTE-NC merupakan perluasan atau generalisasi dari metode SMOTE yang digunakan untuk menangani kumpulan data campuran kontinu dan nominal. Pada data numerik, SMOTE akan bekerja dengan cara mencari *k-nearest neighbors* untuk setiap data dalam kelas minoritas menggunakan jarak *euclidean* (Clawla, 2002).

Misalkan terdapat data yang berdimensi q terdiri dari p variable kontinu dan c variabel nominal, sehingga $q = p + c$

Vektor data minoritas:

$$\mathbf{X}_i = [X_1, X_2, \dots, X_p | X_{p+1}, \dots, X_q]$$

Vektor data tetangga:

$$\mathbf{Z}_i = [Z_1, Z_2, \dots, Z_q | Z_{p+1}, \dots, Z_q]$$

dengan:

$X_i, Z_i \in \mathbb{R}$ untuk $i = 1, \dots, p$ (fitur kontinu)

$X_i, Z_i \in \text{kategori}$ untuk $i = p + 1, \dots, q$ (nominal)

Untuk menghitung kedekatan antara dua vektor pada bagian kontinu digunakan rumus Euclidean sebagai berikut:

$$d_{num}(\mathbf{X}, \mathbf{Z}) = \sqrt{\sum_{i=1}^p (X_i - Z_i)^2} \quad (2)$$

Pembangkitan data buatan yang berskala kontinu berbeda dengan nominal. Data numerik diukur jarak kedekatannya dengan jarak Euclidean sedangkan data kategorik lebih sederhana yaitu dengan nilai modus. Perhitungan jarak antar contoh kelas minor yang peubahnya berskala nominal dilakukan dengan rumus Value Difference Metric (VDM) sebagai berikut (Rais dan Subekti, 2019):

$$d_{cat}(\mathbf{X}, \mathbf{Z}) = w_x w_z \sum_{i=1}^N \delta(x_i, z_i)^r \quad (3)$$

dengan fungsi $\delta(x_i, z_i)$ didefinisikan sebagai:

$$\delta(x_i, z_i) = \begin{cases} 0, & \text{jika } x_i = z_i \\ 1, & \text{jika } x_i \neq z_i \end{cases}$$

dimana:

$d_{cat}(\mathbf{X}, \mathbf{Z})$: jarak antara amatan $\mathbf{X}$ dan $\mathbf{Z}$

$w_x w_z$: bobot masing-masing pengamatan (dapat diabaikan atau diasumsikan 1)

N : banyak peubah penjelas

r : jarak manhattan (bernilai 1) atau jarak euclidean (bernilai 2)

Data sintetis untuk data kontinu akan dihasilkan dengan menggunakan persamaan berikut:

$$X_{syn,num} = X_i + (Z_i - X_i)\gamma, \quad \text{untuk } i = 1, \dots, p \quad (4)$$

dengan:

$X_{syn,num}$: Data sintetis kontinu

X_i : Data dari kelas minoritas

Z_i : Data dari kelas minoritas yang memiliki tetangga terdekat

γ : Angka acak antara 0 dan 1

Data sintetik pada fitur nominal ke- i diambil sebagian modus dari tetangga terdekat sebagai berikut:

$$X_{syn,cat} = \text{modus} (Z_i^{(1)}, Z_i^{(2)}, \dots, Z_i^{(k)}), \text{ untuk } i = P + 1, \dots, q \quad (5)$$

1.6.4 WK-NN

Dalam metode KNN, biasanya diasumsikan bahwa semua tetangga terdekat dari suatu data (kueri) memiliki pengaruh yang sama. Namun, asumsi ini tidak selalu tepat untuk tugas klasifikasi pola. Oleh karena itu, dikembangkan metode KNN dengan pemberian bobot, yang disebut WK-NN (*Weighted K-Nearest Neighbor*). dalam WK-NN, Bobot ω_i untuk tetangga ke- i dari kueri x didefinisikan sebagai berikut (Li and Yang, 2017):

$$\omega_i = \begin{cases} \frac{d(x, n_K) - d(x, n_i)}{d(x, n_K) - d(x, n_1)} & \text{jika } d(x, n_K) \neq d(x, n_1) \\ 1 & \text{jika } d(x, n_K) = d(x, n_1) \end{cases} \quad (6)$$

Dengan:

$d(x, n_i)$: jarak antara titik kueri x dan tetangga ke- i

n_1 : tetangga terdekat pertama

n_K : tetangga terjauh ke- K dari x

1.6.5 Regresi Logistik Biner

Regresi logistik adalah model regresi di mana variabel respon bersifat kategoris dan variabel prediktor yang sesuai dapat bersifat kategoris atau kontinu (Hae et al, 2017). Variabel respon biasanya biner yang hanya mengambil nilai 0 atau 1 meskipun dapat diperluas ke variabel kategoris dengan beberapa kelas dalam model yang lebih kompleks (Lu, 2022). Peubah dependen memiliki dua kategori yaitu BBLR diberikan kode 1 dan berat badan lahir normal dengan kode 0 dengan 7 peubah prediktor yang dapat dilihat pada Tabel 2.1. Model regresi logistik biner dengan p peubah prediktor adalah sebagai berikut (Hosmer et al, 2000):

$$\pi(x_i) = \frac{\exp g(x)}{1 + \exp g(x)} \quad (7)$$

dimana

$\pi(x_i) = P(Y = 1|X = x_i)$ dan $0 \leq \pi(x_i) \leq 1$

Setara dengan transformasi logitnya berikut:

$$\text{logit}[\pi(x_i)] = g(x) = \ln \frac{\pi(x_i)}{1 - \pi(x_i)} = \beta_0 + \beta_1 X_1 + \dots + \beta_p X_p \quad (8)$$

dengan

p : banyaknya peubah prediktor

$X_1, X_2, \dots, X_p$: peubah prediktor

$\beta_1, \beta_2, \dots, \beta_p$: parameter yang tidak diketahui dan perlu diduga nilainya.

1.6.6 Estimasi Parameter

Salah satu metode paling umum untuk mengestimasi parameter dalam regresi logistik biner adalah *maksimum likelihood estimation* atau MLE (Ahmed dan Cheng, 2020). Teknik ini bertujuan untuk menemukan nilai parameter yang memaksimalkan fungsi kemungkinan, yang mewakili probabilitas mengamati data yang diberikan di bawah model (Ulfa, 2020). Fungsi MLE nya adalah sebagai berikut (Pangestika, 2021):

$$L(\beta) = \prod_{i=1}^n \pi(x_i)^{y_i} [1 - \pi(x_i)]^{1-y_i} \quad (9)$$

dengan

i : 1, 2, ... , n

y_i : respon amatan ke- i

$\pi(x_i)$: peluang kejadian ke- i bernilai y_i

Secara matematika rumus diatas lebih mudah dikerjakan dalam bentuk *log-likelihood* yang ditulis sebagai berikut:

$$\ln L(\beta) = \sum_{i=1}^n [y_i \ln \pi(x_i) + (1 - y_i) \ln (1 - \pi(x_i))] \quad (10)$$

Untuk memperoleh nilai parameter β , fungsi log-likelihood tersebut dimaksimumkan dengan menurunkan $\ln L(\beta)$ terhadap masing-masing parameter dan menyamakan hasilnya dengan nol. Namun, berdasarkan Persamaan (15), (16), dan (17), turunan pertama *fungsi log-likelihood* menghasilkan sistem persamaan yang bersifat nonlinear karena mengandung fungsi eksponensial dari parameter β . Akibatnya, tidak tersedia bentuk solusi tertutup secara analitik untuk memperoleh estimasi parameter, maka digunakan pendekatan numerik untuk mendapatkan estimasi parameter. Estimasi parameter β dilakukan melalui prosedur iteratif dengan menggunakan metode *Newton-Raphson*, yang bertujuan untuk memaksimumkan fungsi *log-likelihood* (Agresti, 2002).

1.6.7 Iterasi Newton Raphson

Metode *newton raphson* merupakan metode iterasi untuk menyelesaikan persamaan nonlinear. Berdasarkan iterasi *newton raphson* akan diperoleh turunan kedua dari fungsi log-likelihood terhadap masing-masing parameter. Formulasi iterasi *newton raphson* dengan memulai iterasi $t = 0$ sebagai berikut:

$$\beta_{(t+1)} = \beta_{(t)} - (H_{(t)})^{-1} d_{(t)} \quad (11)$$

dengan:

t : iterasi ke- 0, 1, 2, ... sampai konvergen

$\beta_{(t)}$: Nilai taksiran awal iterasi ke- t

$d_{(t)}$: matriks turunan pertama terhadap masing-masing parameter β

$H_{(t)}$: matriks turunan kedua terhadap masing-masing parameter β

Proses iterasi dengan menggunakan metode *newton raphson* akan berhenti jika terpenuhi kondisi konvergen (Sofiyat et al, 2023).

1.6.8 Pengujian Parameter

Pengujian estimasi parameter model merupakan pengujian yang digunakan untuk menguji signifikansi koefisien β dari model (Pratiwi, 2021). Pengujian ini terdiri dari dua pengujian yaitu secara simultan dan secara parsial sebagai berikut:

1.6.8.1 Uji Simultan

Pengujian secara simultan terhadap parameter pada model dilakukan sebagai upaya untuk memeriksa peranan peubah prediktor yang ada di dalam model. Untuk mengetahui peran seluruh peubah prediktor di dalam model secara simultan digunakan statistik uji- G (Hosmer et al, 2000).

Hipotesis yang di uji adalah:

$$H_0 : \beta_1 = \beta_2 = \dots = \beta_p = 0$$

$$H_1 : \text{paling sedikit ada satu } \beta_i \neq 0, i = 1, 2, \dots, p$$

Statistic uji- G didefinisikan sebagai:

$$G = -2 \ln \frac{L_0}{L_P} \quad (12)$$

Dimana L_0 adalah fungsi kemungkinan maksimum tanpa peubah prediktor, dan L_P merupakan fungsi kemungkinan maksimum dengan p peubah prediktor. Hipotesis nol ditolak jika nilai $G > \chi^2$

1.6.8.2 Uji Parsial

Uji parsial digunakan untuk mengetahui peranan masing-masing variabel prediktor terhadap variabel respon. Uji parameter secara parsial yang digunakan adalah statistik uji- $Wald$ (Hosmer et al, 2000).

Hipotesis yang di uji adalah:

$$H_0 : \beta_i = 0$$

$$H_1 : \beta_i \neq 0, i = 1, 2, \dots, p$$

Statistic uji $Wald$ didefinisikan sebagai:

$$W = \frac{B_i}{\widehat{SE}_{B_i}} \quad (13)$$

dengan

$$i : 1, 2, \dots, p$$

$$B_i : \text{dugaan koefisien parameter ke-}i$$

$$\widehat{SE}_{B_i} : \text{dugaan galat baku dari } B_i$$

Hipotesis nol ditolak jika nilai $W > \chi^2$ atau $p\text{-value} < \alpha$.

1.6.9 Evaluasi Model

Evaluasi model dilakukan dengan menghitung rasio antara jumlah observasi yang diklasifikasikan dengan benar oleh model dan total jumlah observasi (Misna et al., 2018). Pengukuran kinerja model klasifikasi ini melibatkan evaluasi kemungkinan terjadinya misklasifikasi, yang dapat dianalisis menggunakan *Confusion Matrix* (Faisal, 2019). Penentuan klasifikasi ditunjukkan pada Tabel 2.1

Tabel 2.1 *Confussion matrix*

Observation	Predicted		Total
	Positif class (Y=1)	Negative class (Y=0)	
Positive class (Y=1)	True Positive (TP)	False Negative (FN)	TP + FN
Negative class (Y=0)	False Positive (FP)	True Negatif (TN)	FP + TN
Total	TP + FP	FN + TN	<i>n</i>

Sumber (Islahulhaq *et al.*, 2021)

dengan:

TP : Jumlah observasi *class* 1 yang tepat diklasifikasikan sebagai *class* 1.

TN : Jumlah observasi *class* 0 yang tepat diklasifikasikan sebagai *class* 0.

FP : Jumlah observasi *class* 0 yang tidak tepat diklasifikasikan sebagai *class* 1.

FN : Jumlah observasi *class* 1 yang tidak tepat diklasifikasikan sebagai *class* 0

Beberapa indikator dihasilkan dari *confusion matrix* dan dapat digunakan untuk mengukur nilai kinerja model klasifikasi:

1. Akurasi = $[TP+TN / n]$ adalah persentase model klasifikasi yang diprediksi dengan benar
2. Sensitivitas = $[TP/TP+FN]$ adalah persentase observasi kelas positif yang diprediksi dengan benar sebagai kelas positif.
3. Spesifisitas = $[TN / FP + TN]$ adalah persentase observasi kelas negatif yang diprediksi dengan benar sebagai kelas negatif.

1.6.10 Berat Badan Lahir Rendah (BBLR)

Berat badan lahir rendah (BBLR) didefinisikan oleh Organisasi Kesehatan Dunia (WHO) sebagai berat badan saat lahir kurang dari 2.500 gram. BBLR merupakan indikator penting status kesehatan bayi yang memengaruhi kelangsungan hidup, pertumbuhan fisik, dan perkembangan mental bayi (Toru, 2020). Kondisi ini memiliki hubungan signifikan dengan angka kematian neonatal yang menyumbang 60% sampai 80% dari semua kematian neonatal. Prevalensi global BBLR adalah 20 juta pertahun atau 15,5% dari seluruh bayi lahir di dunia (WHO, 2015).

Dampak serius BBLR terhadap bayi baru lahir meliputi risiko gangguan perkembangan kognitif, seperti retardasi mental, serta gangguan kesehatan akibat imaturitas organ. Bayi BBLR yang lahir prematur, misalnya, rentan mengalami asfiksia karena kondisi paru-paru yang belum matang sepenuhnya. Selain itu, sistem imun yang lemah pada bayi BBLR membuat mereka lebih rentan terhadap infeksi, yang dapat menyebabkan kesakitan atau bahkan kematian (Hartiningrum dan Fitriyah, 2018).

Banyak faktor yang menyebabkan bayi lahir dengan berat badan kurang atau berat badan lahir rendah (BBLR), di antaranya adalah Usia kehamilan atau usia gestasi (*gestational age*) yang merupakan lama waktu seorang janin berada dalam rahim terhitung dari Hari Pertama Haid Terakhir (HPHT) sampai ibu melahirkan bayinya. Usia kehamilan memiliki hubungan dengan kejadian berat badan lahir

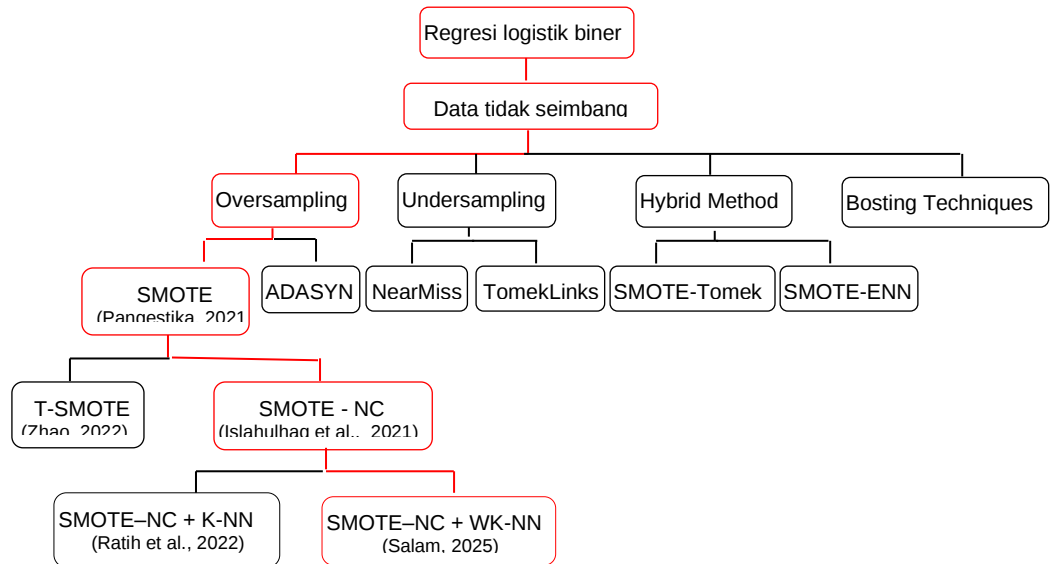
rendah BBLR (Winanti, 2023). Hemoglobin (Hb) juga dikaitkan dengan kejadian BBLR, dimana Hb adalah pigmen pada sel darah merah yang berperan penting dalam mengangkut oksigen dalam tubuh, dan kadar Hb yang terlalu tinggi atau rendah dapat menyebabkan masalah klinis (Khadka, 2021). Kadar hemoglobin ibu hamil yang rendah (di bawah 110 g/L) sering dikaitkan dengan risiko hasil kelahiran yang buruk, termasuk BBLR (Young, 2019). Selain itu, status pekerjaan ibu dapat mempengaruhi kehamilan, beratnya pekerjaan ibu selama kehamilan dapat menimbulkan terjadinya prematuritas dan melahirkan bayi dengan BBLR karena selama hamil ibu tidak dapat beristirahat dan hal tersebut dapat mempengaruhi janin yang dikandungnya (Salawati, 2012). Status pekerjaan ibu hamil digolongkan menjadi dua yaitu bekerja dan tidak bekerja (Ernawati, 2018).

Faktor lain yang berkontribusi terhadap BBLR adalah komplikasi kehamilan yang dimana komplikasi kehamilan merupakan masalah kesehatan yang sering terjadi selama masa kehamilan maupun persalinan. Masalah ini dapat dimulai sejak sebelum kehamilan dan berdampak pada kesehatan ibu serta bayi yang dilahirkan (Annisa and Iwan, 2021). Studi terkini menunjukkan bahwa komplikasi maternal merupakan salah satu kontributor utama komplikasi neonatal. Risiko komplikasi maternal lebih sering diamati pada kehamilan berisiko tinggi, terutama pada wanita usia lanjut. Kondisi seperti hipertensi gestasional, anemia maternal, diabetes gestasional, dan perdarahan antepartum secara signifikan meningkatkan peluang terjadinya hasil neonatal yang buruk. Sebagai contoh, diabetes gestasional diketahui dapat meningkatkan risiko makrosomia dan kebutuhan untuk persalinan caesar. Selain itu, faktor seperti tingkat konsanguinitas yang tinggi, usia ibu lanjut, dan riwayat obstetrik yang buruk juga berkontribusi pada meningkatnya risiko anomali kongenital pada bayi baru lahir, terutama dari wanita usia lanjut. Berat badan lahir rendah juga lebih sering ditemukan pada kasus anemia maternal dan perdarahan antepartum (Bener, 2012). Kemudian, paritas yang tinggi dapat berdampak serius pada kesehatan ibu dan bayi, termasuk risiko melahirkan bayi dengan berat badan lahir rendah (BBLR). Kehamilan dan persalinan yang berulang menyebabkan kerusakan pada pembuluh darah di dinding rahim serta penurunan elastisitas jaringan akibat regangan berulang selama kehamilan. Kondisi ini meningkatkan kemungkinan terjadinya kelainan letak, kelainan pertumbuhan plasenta, dan gangguan pertumbuhan janin yang berujung pada BBLR (Nurseha and Annisa, 2017). Secara statistik, paritas memiliki hubungan signifikan dengan kejadian BBLR, yang dapat dicegah melalui perawatan prenatal yang optimal dan program intervensi yang tepat (Kachapati, 2022). Selain itu, usia ibu memengaruhi berat lahir neonatal, dengan tingkat berat lahir rendah yang jauh lebih tinggi di kalangan ibu remaja (Sekhavat, 2023).

1.6.11 Kerangka Konseptual

Kerangka konsep penelitian ini menggambarkan hubungan antar konsep yang digunakan sebagai dasar teori. Gambar 1 menunjukkan alur metode regresi logistik biner untuk memodelkan hubungan antara variabel prediktor dan respon kategorikal dua kelas. Ketidakseimbangan data, yang umum terjadi, dapat menyebabkan model

bias terhadap kelas mayoritas. Untuk mengatasinya, digunakan metode. SMOTE-NC dengan WKNN memungkinkan penanganan data dengan variabel nominal dan kontinu sekaligus, sehingga meningkatkan kualitas klasifikasi. Kerangka penelitian ini disajikan pada Gambar 1.



Gambar 1. Kerangka Konseptual

BAB II

METODE PENELITIAN

2.1 Sumber Data

Data yang digunakan dalam penelitian ini merupakan data sekunder mengenai populasi berat badan lahir bayi selama periode Tahun 2022. Data diperoleh dari Rekam Medik Puskesmas di Kabupaten Buton Tengah wilayah Kecamatan Mawasangka dan wilayah Kecamatan Gu yang kemudian ditampilkan pada Lampiran 1.

2.2 Variabel Penelitian

berikut variabel-variabel yang menjadi fokus pada penelitian ini:

Tabel 2.1 Variabel Penelitian

Variabel	Kategori	Definisi Operasional	Skala
Status Berat Badan Lahir (Y)	0 : Tidak BBLR	Berat badan bayi kelahiran lebih dari sama dengan 2500 gram	Nominal
	1 : BBLR	Berat badan bayi kelahiran kurang dari 2500 gram	
Umur Kehamilan (X ₁)	dalam minggu	Umur kehamilan saat melahirkan (minggu)	Rasio
Kadar Hb (X ₂)	0 : normal	Kadar Hb lebih dari sama dengan 11 gr/dl	Nominal
	1 : anemia	Kadar Hb kurang dari 11 gr/dl	
Status Pekerjaan Ibu (X ₃)	0 : Tidak Bekerja	Selama kehamilan ibu tidak bekerja (ibu rumah tangga) atau melakukan aktivitas berat	Nominal
	1 : Bekerja	Selama kehamilan ibu memiliki pekerjaan seperti guru honorer, PNS, swasta maupun wiraswasta	
Komplikasi Kehamilan (X ₄)	0 : Tidak	Tidak mengalami infeksi dalam kehamilan	Nominal
	1 : Ya	Mengalami infeksi dalam kehamilan seperti ancaman persalinan prematur, DBD, malaria dan lain sebagainya	
Paritas (X ₅)	Jumlah anak	Jumlah anak yang telah dilahirkan ibu baik lahir hidup maupun lahir mati	Ratio
Usia Ibu (X ₆)	dalam tahun	Usia ibu saat persalinan	Ratio

2.3 Metode Analisis Data

Langkah-langkah yang dilakukan dalam analisis data untuk mencapai tujuan penelitian ini adalah:

1. Menginput dan melakukan analisis statistika deskriptif.
2. Melakukan uji independensi untuk memeriksa signifikansi tiap variabel prediktor dengan variabel respon menggunakan uji *chi square*.
3. Membagi data menjadi data *training* dan data *testing* dengan perbandingan 80% dan 20%.
4. Melakukan SMOTE-NC dengan WK-NN dengan oversampling 600% dan K=5, proses ini melibatkan langkah-langkah berikut:
 - a. Mengidentifikasi jenis variabel pada data minoritas.
 - b. Menghitung jarak antar data minoritas. Untuk variabel kontinu digunakan jarak Euclidean seperti pada Persamaan (2). Sedangkan, variabel nominal digunakan Value Difference Metric (VDM) seperti pada Persamaan (3).
 - c. Menggabungkan jarak kontinu dan nominal. Total jarak diperoleh dengan menjumlahkan kedua jenis jarak. Jika terdapat perbedaan pada variabel nominal, median dari standar deviasi variabel kontinu digunakan sebagai penyesuaian.
 - d. Menentukan tetangga terdekat. Dipilih berdasarkan jarak terkecil dari hasil gabungan jarak numerik dan kategorik.
 - e. Mengaplikasikan metode WK-NN. Bobot tiap tetangga dalam pembentukan data sintetis dihitung berdasarkan Persamaan (6) yang mempertimbangkan seberapa dekat masing-masing tetangga dengan titik pusat.
 - f. Membuat data sintetis untuk variabel kontinu. Nilai data baru dihitung dengan interpolasi antara data asli dan tetangganya menggunakan Persamaan (4). Kemudian, untuk membuat data sintetis untuk variabel kategorik, ditentukan berdasarkan modus dari tetangga terdekat, seperti pada Persamaan (5).
 - g. Menggabungkan data sintetis dengan data asli. Hasil dari proses SMOTE-NC dengan WK-NN ditambahkan ke dalam data training untuk digunakan dalam tahap modeling selanjutnya.
5. Melakukan estimasi parameter regresi logistik biner setelah dilakukan SMOTE-NC dengan WK-NN untuk menangani data tidak seimbang. Proses estimasinya sebagai berikut:
 - a. Menyusun bentuk model regresi logistik biner. Model logistik biner dibentuk menggunakan fungsi logistik sebagaimana pada Persamaan (7), dengan bentuk logit seperti pada Persamaan (8).
 - b. Menyusun fungsi likelihood dan log-likelihood. Fungsi likelihood dinyatakan seperti pada Persamaan (9), dan log-likelihood-nya dinyatakan dalam bentuk yang lebih mudah dioperasikan seperti pada Persamaan (10) hingga Persamaan (14).
 - c. Menghitung turunan pertama fungsi log-likelihood. Turunan pertama terhadap parameter β_0 dan β_j dihitung seperti pada Persamaan (15) dan Persamaan (16), kemudian dibentuk menjadi vektor gradien seperti pada Persamaan (17).

- d. Menghitung turunan kedua (Hessian matrix). Turunan kedua terhadap β_0 , campuran β_0 dan β_j , serta campuran β_j dan β_m dihitung sesuai dengan Persamaan (18), (19), dan (20), yang membentuk matriks Hessian.
 - e. Melakukan iterasi *newton raphson* untuk memperoleh estimasi parameter. Proses iterasi dilakukan menggunakan prosedur *newton raphson* seperti pada Persamaan (11), dengan memperhitungkan turunan pertama dan kedua.
 - f. Mendapatkan nilai parameter taksiran maksimum likelihood. Setelah iterasi mencapai konvergensi, diperoleh nilai taksiran parameter seperti pada Persamaan (21), yang akan digunakan dalam interpretasi model dan prediksi klasifikasi.
6. Melakukan pengujian signifikansi parameter secara simultan menggunakan Uji-G berdasarkan Persamaan (12)
 7. Menguji signifikansi parameter secara parsial menggunakan Uji-Wald berdasarkan Persamaan (13)
 8. Melakukan estimasi parameter ulang untuk mendapatkan model yang terbaik berdasarkan hasil pengujian parameter secara parsial
 9. Evaluasi prosedur klasifikasi untuk mengukur kinerja model klasifikasi.